Segment-based Acoustic Models for Continuous Speech Recognition

Progress Report: April – June 1993

submitted to
Office of Naval Research
and
Advanced Research Projects Administration
8 July 1993

by
Boston University
Boston, Massachusetts 02215

Principal Investigators

Dr. Mari Ostendorf
Assistant Professor of ECS Engineering, Boston University
Telephone: (617) 353-5430

Dr. J. Robin Rohlicek
Scientist, BBN Inc.
Telephone: (617) 873-3894

Administrative Contact

Maureen Rogers, Awards Manager
Office of Sponsored Programs
Telephone: (617) 353-4365

Executive Summary

This research aims to develop new and more accurate stochastic models for speaker-independent continuous speech recognition, by extending previous work in segment-based modeling and by introducing a new hierarchical approach to representing intra-utterance statistical dependencies. These techniques, which are more costly than traditional approaches because of the large search space associated with higher order models, are made feasible through rescoring a set of HMM-generated N-best sentence hypotheses. We expect these different modeling techniques to result in improved recognition performance over that achieved by current systems, which handle only frame-based observations and assume that these observations are independent given an underlying state sequence.

In the fourth quarter of the project, we have: (1) ported our recognition system to the Wall Street Journal task, a standard task in the ARPA community; (2) developed an initial dependency-tree model of intra-utterance observation correlation; and (3) implemented baseline language model estimation software. Our initial results on the Wall Street Journal task are quite good, representing improved performance over most HMM systems reporting on the November 1992 5k vocabulary test set.

Contents

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 April 1993 – 30 June 1993

1 Productivity Measures

- Refereed papers submitted but not yet published: 0
- Refereed papers published: 0
- Unrefereed reports and articles: 1
- Books or parts thereof submitted but not yet published: 0
- Books or parts thereof published: 0
- Patents filed but not yet granted: 0
- Patents granted (include software copyrights): 0
- Invited presentations: 0
- Contributed presentations: 0
- Honors received:
- Prizes or awards received: 0
- Promotions obtained: 0
- Graduate students supported $\geq$ 25% of full time: 3
- Post-docs supported $\geq$ 25% of full time: 0
- Minorities supported: 1 woman

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 April 1993 – 30 June 1993

2 Summary of Technical Progress

Introduction and Background

In this work, we are interested in the problem of large vocabulary, speaker-independent continuous speech recognition, and primarily in the acoustic modeling component of this problem. In developing acoustic models for speech recognition, we have conflicting goals. On one hand, the models should be robust to inter- and intra-speaker variability, to the use of a different vocabulary in recognition than in training, and to the effects of moderately noisy environments. In order to accomplish this, we need to model gross features and global trends. On the other hand, the models must be sensitive and detailed enough to detect fine acoustic differences between similar words in a large vocabulary task. To answer these opposing demands requires improvements in acoustic modeling at several levels. New signal processing or feature extraction techniques can provide more robust features as well as capture more acoustic detail. Advances in segment-based modeling can be used to take advantage of spectral dynamics and segment-based features in classification. Finally, a new structural context is needed to model the intra-utterance dependence across phonemes.

This project addresses some of these modeling problems, specifically advances in segment-based modeling and development of a new formalism for representing inter-model dependencies. The research strategy includes three main thrusts. First, speech recognition is implemented under the N-best rescoring paradigm [1], in which the BBN Byblos system is used to constrain the stochastic segment model (SSM) search space by providing the top N sentence hypotheses. This paradigm facilitates research on the segment model through reducing development costs, and provides a modular framework for technology transfer that has already enabled us to advance state-of-the-art recognition performance through collaboration with BBN. Second, we are working on improved segment modeling at the phoneme level [2, 3, 4] by developing new techniques for robust context modeling, mechanisms for effectively incorporating segmental features, and models of within-segment dependence of frame-based features. Lastly, we plan to investigate hierarchical structures for representing the intra-utterance dependency of phonetic models in order to capture speaker-dependent and session-dependent effects within the context of a speaker-independent model. Additionally, we have expanded the scope of our work to include some language modeling, recognizing that

higher-order models of correlation can extend to this domain as well.

Summary of Recent Technical Results

In much of the first year of the project, we focused on improving the performance of the basic SSM word recognition system. In brief, the accomplishments of that period included: improvements to the N-Best rescoring technique by introducing score normalization; development of a method for clustering contexts to provide robust context-dependent model parameter estimates [5]; extensions to the classification and segmentation scoring formalism to handle context-dependent models with long-range acoustic features; extension of the two level segment/microsegment formalism and assessment of trade-offs in mixture vs. trajectory modeling [6]; development and assessment of automatically generated multiple-pronunciation word networks; and investigation of the use of tied mixtures in the segment model [7].

The research efforts during this quarter have focused on porting the BU recognition system to the Wall Street Journal (WSJ) domain, and beginning development of models (both acoustic and language) that will capture higher level dependencies in speech. In particular, we have:

Ported the SSM word recognition system to the Wall Street Journal task domain: The effort to port our recognition system to the WSJ domain involved modifying functions to maintain compatibility with BBN, modifying I/O formats to handle the new dictionary, and porting three variations of the SSM trainer and recognizer. On the November 1992 5k vocabulary test set, using the standard bigram language model, we achieved the following results:

	% Word Error	
SSM System	SSM	SSM+HMM
Baseline	8.1	7.5
Clustered covariances	8.1	7.6
Tied mixtures	9.2	8.1

which can be compared to the BBN HMM result of 8.7% [8]. These experiments confirmed previous results on the Resource Management (RM) corpus, that covariance clustering significantly reduces storage (by a factor of ten for covariance parameters) without any reduction in recognition performance. In fact, the clustered system worked slightly better on our development test set. Unfortunately, previous results on RM showing improvements with tied mixtures were not confirmed in the WSJ experiments, and we intend to explore more specific regions of tying that have provided performance gains for other sites.

Developed an initial dependence-tree model of intra-utterance observation correlation: An important goal of this project is the development of a hierarchical model of intra-utterance correlation of phone observations. Our initial efforts in this area have been to extend the work of Chow and

Liu on dependence trees [9] from discrete models to Gauss-Markov dependencies. We are currently implementing the algorithm to find the minimal spanning tree, which is based on a mutual information measure assuming a joint Gaussian model of phone observation vectors. In order to quickly assess different models of dependence without the high cost of building a full word recognition system, we plan to initially compare prediction errors for different models within the context of the TIMIT corpus.

Implemented baseline language model estimation software for the WSJ task: Motivated by the realization that inter- and intra-utterance correlation can be modeled at the language as well as acoustic level, we have begun an effort in dynamic language modeling. As a first step in this project, we have implemented the back-off algorithm for estimating n-gram language models [10] and an efficient storage mechanism. In experiments on the 5k vocabulary, we have duplicated the bigram perplexity measures reported in the literature, which we will use as a baseline against which to measure improvements due to better modeling. We are currently developing a mixture n-gram language model to better represent the topic-dependent structure of language.

Future Goals

Based on the results of the past year and our original goals for the project, we have set the following goals for the next six months: (1) modify BBN's current N-best search algorithm to provide lattice outputs for rescoring with the SSM; (2) further develop the hierarchical model formalism and assess the trade-offs between linear and non-linear models of dependence; (3) implement a dynamic language model and assess in the WSJ task domain; and (4) investigate unsupervised adaptation in the WSJ task domain.

References

[1] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, J. R. Rohlicek, "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses," *Proc. of the DARPA Workshop on Speech and Natural Language*, pp. 83-87, February 1991.

[2] M. Ostendorf and S. Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. Acoustics Speech and Signal Processing, Dec. 1989.*

[3] S. Roucos, M. Ostendorf, H. Gish, and A. Derr, "Stochastic Segment Modeling Using the Estimate-Maximize Algorithm," *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 127–130, New York, New York, April 1988.

[4] M. Ostendorf, A. Kannan, O. Kimball and J. R. Rohlicek, "Continuous Word Recognition Based on the Stochastic Segment Model," *Proceedings of the 1992 DARPA Workshop on Artificial Neural Networks and Continuous Speech Recognition*, September 1992.

[5] A. Kannan, M. Ostendorf and J. R. Rohlicek, "Maximum Likelihood Clustering of Gaussians for Speech Recognition," in review.

[6] A. Kannan and M. Ostendorf, "A Comparison of Trajectory and Mixture Modeling in Segment-based Word Recognition," *Proc. of the Inter. Conf. on Acoust., Speech and Signal Processing*, pp. II327-330, April 1993.

[7] O. Kimball and M. Ostendorf, "On the Use of Tied Mixture Distributions," *Proceedings of the ARPA Workshop on Human Language Technology*, 1993.

[8] D. S. Pallett, J. G. Fiscus, W. M. Fisher, and J. S. Garofolo, "Benchmark Tests for the DARPA Spoken Language Program, *ARPA Workshop on Human Language Technology*, March 21-24, Plainsboro, NJ.

[9] C. K. Chow and C. N. Liu, "Approximating Discrete Probability Distributions with Dependence Trees," *IEEE Transactions on Information Theory*, Vol. IT-14, No. 3, May 1968, pp. 463-467.

[10] S. M. Katz, "Estimation of probabilities from sparse data for the LM component of a Speech Recognizer," *IEEE Trans Vol. ASSP-35*, No. 3, March 1987.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 April 1993 – 30 June 1993

3 Publications and Presentations

Papers appearing during the reporting period include one conference paper, a copy of which is included with the report.

- "A Comparison of Trajectory and Mixture Modeling in Segment-based Word Recognition," A. Kannan and M. Ostendorf, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* pp. II327-330, April 1993.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 April 1993 - 30 June 1993

4 Transitions and DoD Interactions

This grant includes a subcontract to BBN, and the research results and software is available to them. Thus far, we have collaborated with BBN by combining the Byblos system with the SSM in N-Best sentence rescoring to obtain improved recognition performance, and we have provided BBN with papers and technical reports to facilitate sharing of algorithmic improvements. On their part, BBN has been very helpful to us in our WSJ porting efforts, providing us with WSJ data and consulting on format changes.

The recognition system that has been developed under the support of this grant and of a joint NSF-DARPA grant (NSF # IRI-8902124) is currently being used for automatically obtaining good quality phonetic alignments for a corpus of radio news speech under development at Boston University. The alignment effort is supported by the Linguistic Data Consortium, through a grant that allowed us to add cross-word phonological rules to the segmentation software.

Principal Investigator Name: Mari Ostendorf
PI Institution: Boston University
PI Phone Number: 617-353-5430
PI E-mail Address: mo@raven.bu.edu
Grant or Contract Title: Segment-Based Acoustic Models for Continuous Speech Recognition
Grant or Contract Number: ONR-N00014-92-J-1778
Reporting Period: 1 April 1993 – 30 June 1993

5 Software and Hardware Prototypes

Our research has required the development and refinement of software systems for parameter estimation and recognition search, which are implemented in C or C++ and run on Sun Sparc workstations. No commercialization is planned at this time.

A COMPARISON OF TRAJECTORY AND MIXTURE MODELING IN SEGMENT-BASED WORD RECOGNITION

Ashvin Kannan *Mari Ostendorf*

Electrical, Computer and Systems Engineering
Boston University
Boston, MA 02215, USA

ABSTRACT

This paper presents a mechanism for implementing mixtures at a phone-subsegment (microsegment) level for continuous word recognition based on the Stochastic Segment Model (SSM). We investigate the issues that are involved in trade-offs between trajectory and mixture modeling in segment-based word recognition. Experimental results are reported on DARPA's speaker-independent Resource Management corpus.

1. INTRODUCTION

In earlier work, the Stochastic Segment Model (SSM) [1, 2] has been shown to be a viable alternative to the Hidden Markov Model (HMM) for representing variable-duration phones. The SSM provides a joint Gaussian model for a sequence of observations. Assuming each segment generates an observation sequence of random length, the model for a phone consists of 1) a family of joint density functions (one for every observation length), and 2) a collection of mappings that specify the particular density function for a given observation length. Typically, the model assumes that segments are described by a fixed-length sequence of locally time-invariant regions (or regions of tied distribution parameters). A deterministic mapping specifies which region corresponds to each observation vector.

A framework has recently been proposed for modeling speech at the microsegment level (a unit smaller than a phone segment) [3], in addition to the segment and frame level. Initial experiments with context-independent (CI) phone classification suggested that microsegment models provided a significant gain over the standard SSM when both models assumed conditional independence of frames given the phone segmentation. In this paper, we modify the microsegment framework for word recognition, extend it to context-dependent (CD) modeling using mixture distributions, and investigate the trade-offs of using more distributions per microsegment (model length) versus more mixture components. We present experimental results on the Resource Management task, and conclude with a discussion of our results and possible future work.

2. MICROSEGMENT FRAMEWORK

The framework consists of two levels: the upper level represented by phones and the lower level represented by microsegments (MS). Each phone-length segment is divided into a fixed number of MS-sized regions. A region is characterized by a set of MS models, each an independent-frame SSM with a fixed number of distributions (multivariate Gaussians) representing a variable-length sequence of frame-level observations. The number of distributions (or MS model length) may vary across regions but is constant for different MS models representing the same region. We use a deterministic linear warping to obtain the MS-level segmentation within a phone segment, since dynamic segmentation did not lead to improved performance [3] and is much more expensive.

The sequence of MS labels can be modeled using a variety of techniques. In [3], the sequence is modeled as a first-order Markov chain, an assumption that was also used in this work for CI models. For CD models, however, the computation was too costly given the minimal benefit over independent MS regions. Consequently, for the CD MS system, we represent only marginal probabilities of the microsegment regions, which is equivalent to a mixture distribution at the microsegment level. Thus the probability of an observed segment Y given phone α is defined as:

$$p(Y|\alpha) = \prod_i \sum_{a_i} p(Y_i|a_i, \alpha) p(a_i|\alpha) \qquad (1)$$

where Y_i and a_i represent observations and MS labels respectively for MS region i. The components of the MS mixture are MS models $p(Y_i|a_i)$ and the probabilities $p(a_i|\alpha)$ which serve as mixture weights. In earlier work [3], it was found that tied-mixtures (sharing the mixture components across all phones) produced poor results, so tied mixtures were not explored here.

We implemented three MS systems and compared their performance with the 8-distribution long SSM. The (3,2,3) system used three MS regions in a segment with 3 distributions in the first and last MS region and

0-7803-0946-4/93$3.00 1993IEEE

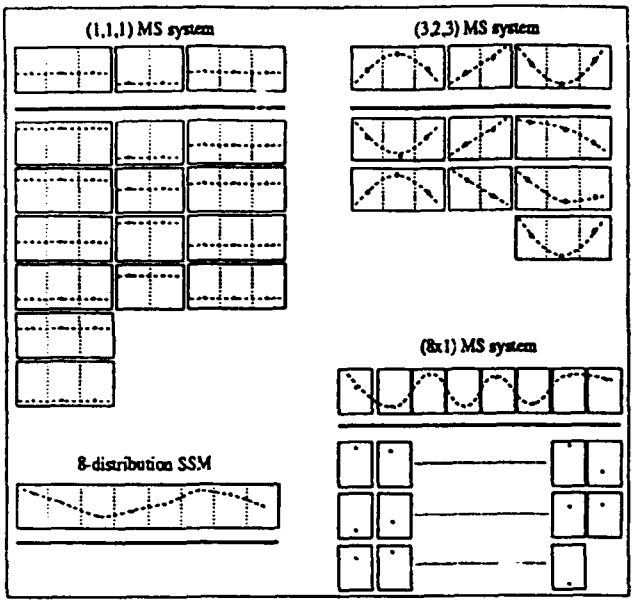

Figure 1: Trajectory assumptions (illustrated for one feature) for the SSM and MS systems. Clockwise from top-left, (1,1,1), (3,2,3), (8x1) MS systems and 8-distribution SSM. Mixture components (when present) are shown below the solid line.

2 distributions in the middle MS region. The (1,1,1) system used three regions with one distribution length each, and the (8x1) system used 8 regions each one distribution long. These systems make different assumptions about the modeling of trajectories of features of speech. The (3,2,3) system assumes that trajectories move within a region, while the (1,1,1) system assumes trajectories are fixed within a region but has more mixture components. The (8x1) system assumes no restriction on the trajectories, and has the same form as 8-distribution SSM except that the distributions are mixtures. These trajectory assumptions are schematically illustrated for one feature in Figure 1.

3. RECOGNITION

Implementation of the recognition search involves a dynamic programming or Viterbi search at the segment level, as for other SSM systems. For the microsegment framework, the difference from the standard SSM is the computation of the probability of a segment for a hypothesized phone label, which can be implemented either as a mixture distribution (as in Equation 1) or approximated by finding the most likely MS sequence. Both methods were investigated here.

The segment probability computation based on the dominant mixture components was investigated to reduce recognition search costs. Under this mode, the search jointly finds the most probable phone and MS sequence, replacing the probability $p(Y|\alpha)$ by the approximation

$$p(Y|\alpha) \approx \max_A p(Y, A|\alpha) = \prod_i \max_{a_i} p(a_i|\alpha) p(Y_i|a_i, \alpha),$$

where A represents an MS label sequence for the phone α. (Note that, for the Markov MS label sequence assumption, $p(a_i|\alpha)$ is replaced by $p(a_i|a_{i-1}, \alpha)$ and a MS-level dynamic programming search is needed.) As we allow for a variable number of microsegment components per region, choosing the dominant component of the mixture results in the grammar introducing differing penalties on phones with different numbers of mixture components. Therefore, the grammar is used in determining the best MS sequence but left out from the segment acoustic probability, i.e.,

$$p(Y|\alpha) \approx p(Y|\hat{A}, \alpha) \approx \prod_i p(Y_i|\hat{a}_i, \alpha), \qquad (2)$$

and this algorithm is what is referred to here as "Viterbi" recognition. In experiments, it was observed that the grammar probabilities had no effect on recognition performance.

4. ESTIMATION OF MS PARAMETERS

Estimation of MS parameters involves estimating means and covariances of their associated Gaussians and the grammar probabilities for the MS units. We first describe the basic procedure and then describe extensions to context modeling.

4.1. Basic procedure

Since the microsegments do not correspond to any linguistic unit, we need to automatically determine and label them in the training database. Training of MS parameters involves the following steps:

1. With the phone segmentation fixed, find initial estimates of MS models –
   (a) Use binary divisive clustering on data to get initial means and partitions.
   (b) Use K-means to improve partitions and define microsegments labels.
   (c) Find maximum-likelihood estimates of mixture components with the partitions found in 1 (b).
2. Use segmental K-means to iteratively improve mixture component parameter estimates –
   (a) Segment speech with current MS parameters.
   (b) Find maximum-likelihood estimates of the MS parameters with the new segmentation.

These steps are described in more detail below.

Initialization

Each MS region is initialized independently of other regions. For each m-distribution long MS region, an n-ary tree with one node for each phone is specified. Each node consists of all the observations from the training set that map to this particular phone and MS region according to the deterministic linear warping. To split a node in step 1 (a), K-means clustering with K=2 is performed at the microsegment level (the mean of a cell is of dimension $m \times k$, where k is the dimension of the feature vector), using a Mahalanobis distance and a linear time warping to map observed frames to regions in the microsegment. A greedy-growing algorithm is used to split the node with the maximum reduction of node distortion. The reduction of node distortion is the difference between the total distortion of the parent node and the sum of the total distortions of the two child nodes, where the distortion of a node is defined as the sum of length-normalized microsegment distances from the mean.

The number of terminal nodes is constrained so that the number of free parameters are comparable across experiments. Specifically, for the CI experiments the number of terminal nodes is equal to three times the number of initial nodes, resulting in three times as many parameters as that used in the CI 8-distribution SSM. After the tree has been fully grown, K-means clustering is performed within each phone sub-tree, to obtain better estimates (Step 1(b)). The resulting clusters define the phone-dependent MS alphabet, referred to here as the CI MS alphabet. The means and covariances of the observations in the terminal nodes are the initial estimates for the CI MS models.

Iterative segmentation/re-estimation

Once initial estimates for the MS models are available, a segmental K-means procedure is used to obtain better estimates. This involves iterating between segmenting speech into microsegments using the current MS parameters and finding new maximum-likelihood estimates for the MS models from the segmented speech.

Bigram and marginal probabilities of the MS labels ($p(a_i|a_{i-1},\alpha)$ and $p(a_i|\alpha)$, respectively) are given by the relative frequencies observed after each segmentation pass. The bigram probabilities, which are used only for experiments with the 3-region CI MS alphabet, are smoothed with the *a priori* probabilities. During recognition it was observed that the grammar score is two orders of magnitude smaller than the acoustic score of the microsegments and its exclusion does not affect recognition performance with the Viterbi search.

4.2. Context Modeling

Context modeling with microsegments is not practical with equivalents of "diphones" or "triphones", since the alphabet size is much larger than that for phones. Instead we define context classes by the collection of triphones at the terminal nodes of the context tree grown using binary divisive clustering as in [4], but with the generalized likelihood ratio distance measure [5, 6].

Once we define context classes to use, we can model context using microsegments in different ways and two schemes were evaluated. First, we can retain the CI MS alphabet[1] and estimate models for these labels conditioned on the context classes In this case, we estimate CD models from the MS observations that are assigned a CI label according to the training segmentation and also correspond to the specific context class. Alternatively, we can incorporate information of the context classes in the MS initialization process and obtain a CD MS alphabet. In this case, the MS tree growing procedure is modified to start with a node for each context class for each phone, with observations arising from that specific context class and that MS region. The tree is grown until we have the desired number of terminal nodes. The rest of the procedure is analogous to the estimation of CI MS acoustic models.

The current approach to estimating the CD MS alphabet results in many fewer free parameters than the context-dependent system based on the CI MS alphabet. In order to compare systems with similar numbers of free parameters, the MS tree growing algorithm was modified such that the tree is grown beyond the first-level "terminal" nodes (called "covariance nodes" and having at least 250 observations to estimate a full covariance) to a second-level set of terminal nodes ("mean nodes") based on a lower threshold, i.e. 50 observations. The mean nodes now constitute an "extended" alphabet and share the covariance of their parent covariance node.

5. EXPERIMENTAL CONDITIONS

Word recognition with the MS-based SSM is performed using the N-best rescoring formalism [2] on DARPA's Resource Management speaker-independent corpus with the word-pair grammar. Gender-dependent MS models are trained on the SI-109, 3990 utterance set. The systems use frame-based observations that include 14 mel-warped cepstra and their first differences, plus the first difference of log energy.

Development was performed on the February 1989 test set and results are also reported on the October 1989 test set. The experimental results for the different systems using Viterbi recognition are shown in Table 1. For the CI MS systems, we see that it is better to have more mixture components than mixtures

[1] For context-modeling experiments, "CI MS alphabet" refers to using the MS labels that were produced from the CI MS tree. In the strict sense, this is not really CI as during re-estimation of the models we use context-dependent variants of these labels. However, we use this nomenclature to differentiate this from the "CD MS alphabet" that is introduced later.

MS System	Average Word Error (%)		
	(8x1)	(3,2,3)	(1,1,1)
Context-independent	7.8	7.6	7.3
CD with CD MS alph.	-	6.3	6.5
CD with CI MS alph.	-	5.8	6.1

Table 1: Performance of the MS systems using Viterbi recognition on the February 89 test set. The 8-distribution SSM achieves 8.9% and 4.8% word error for CI and CD models respectively on this test set.

of sequences since the (1,1,1) system has the best CI performance. On the other hand, for CD systems, it is more important to model the trajectory, since the (3,2,3) system outperforms the (1,1,1) system. In addition, the 8-distribution CD SSM, which does not use mixtures and models the trajectories at the segment rather than the MS level has the best performance.

The initial experiments showed that the CI MS alphabet gave better performance than the CD MS alphabet. However, these systems were not comparable because of differences in the number of free parameters, so further experiments were conducted with the extended CD MS alphabet and the (3,2,3) case using a comparable number of means in both cases. The best CD alphabet system in this case had a maximum of five mean nodes per covariance node. Viterbi recognition for this system resulted in 6.1% word error for the February 89 task while mixture recognition resulted in 5.8%, which was also achieved with the CI alphabet. However, on an independent test set (October 89), the CD alphabet system performed poorly with both Viterbi and mixture recognition. Thus, we conclude that the CI alphabet gives more robust CD models.

We evaluated the best case MS systems, CI (1,1,1) system and the CD (3,2,3) system based on the CI alphabet, on the October 89 test set. The recognition performances were 7.0% and 6.0% respectively. The performance of a comparable 8-distribution SSM on this test set were 8.7% and 4.7% for CI and triphone systems respectively. (Lower error rates have been obtained with more recent system modifications.) Although the microsegment formalism does not yield performance improvements for the CD SSM, it does seem to be preferable in combination with the HMM scores from BBN's Byblos using the N-best rescoring formalism: the word error rate drops to 3.1% on the Oct89 test set from 3.4% for the 8-distribution triphone SSM. For comparison, the Byblos HMM error rate is 3.8%.

6. CONCLUSIONS

In summary, we have described a mechanism for implementing mixtures at a microsegment level and investigated trajectory assumptions for the acoustic modeling for continuous word recognition. Our results suggest that there is a trade-off in using mixture models and trajectory models, associated with the level of detail of the modeling unit (e.g., CI vs. CD), although some level of trajectory constraints is useful even for CI models. The results support the use of whole segment models in the context-dependent case, and microsegment-level (and possibly segment-level) mixtures rather than frame-level mixtures.

In the "mixture" implementation of recognition, we used MS models which were not trained using a "true" mixture procedure, but with the segmentation produced by the dominant component of the best scoring mixture, i.e., with a Viterbi-style training. Performing mixture training may improve performance further. Another possible extension is to further investigate the use of tied microsegment mixtures. Although previous work suggested that tied MS mixtures were not useful, these results were based on region-dependent mixtures, which we have since found are not robust in recent experiments with frame-based mixtures in the SSM.

ACKNOWLEDGMENTS

The authors gratefully acknowledge BBN Inc. for their help in providing the N-best sentence hypotheses. We thank J. Robin Rohlicek of BBN and Vassilios Digalakis of SRI for useful discussions. This research was jointly funded by NSF and DARPA under NSF grant number IRI-8902124, and by DARPA and ONR under ONR grant number N00014-92-J-1778.

REFERENCES

[1] M. Ostendorf and S. Roukos, "A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition," *IEEE Trans. on Acoust., Speech and Signal Processing*, pp. 1857–1869, December 1989.

[2] M. Ostendorf, A. Kannan, O. Kimball and J. R. Rohlicek, "Continuous Word Recognition Based on the Stochastic Segment Model," *Proceedings of the DARPA Workshop on Continuous Speech Recognition*, September 1992.

[3] V. Digalakis, *Segment-Based Stochastic Models of Spectral Dynamics for Continuous Speech Recognition*, Boston University Ph.D. Dissertation, 1992.

[4] K.-F. Lee, S. Hayamizu, H.-W. Hon, C. Huang, J. Swartz, R. Weide, "Allophone Clustering for Continuous Speech Recognition," *Proceedings IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 749–752, April 1990.

[5] H. Gish, M. Siu, R. Rohlicek, "Segregation of Speakers for Speech Recognition and Speaker Identification", *Proceedings IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 873–876, May 1991.

[6] A. Kannan, *Robust Estimation of Stochastic Segment Models for Word Recognition*, Boston University MS Thesis, 1992.